

The creation of the Belgian Bilingual Bi-encoded Thesaurus (3BT)

The starting point for development of the 3BT was the original Dutch version of the Amsterdam Thesaurus, in which the Ghent University primary care group participated by giving feed back in the development of it.

Official presentation and publication of this Dutch version: Bilthoven May 2003.

Later on this Amsterdam Thesaurus (AT) has been translated and published in English, and is now included on Cd-rom in the book ICPC-2-R.

Basic comments for starting up our changes to the AT:

1. The AT does not contain text labels suitable for immediate use in the EPR: only the ICPC/ICD codes or the ICD and ICPC titles as text are ready for implementation in the EPR.
2. The AT used as basic access the ICD system: the search text to find the appropriate code links to ICD and all of the ICPC codes for that ICD code, even if this link to ICPC for that search text is not correct. The AT managed only the correct ICD link, but is not "ICPC oriented".
3. All of the search terms in the AT are double or even triple and more: all the search terms in different order of the included words have been added as separate files. This is not useful in an actual EPR, because the order of the used search terms has no importance in finding the right results in an actual computer system. (The AT has been based on a DOS version of EPR system).
4. The problem of 'aetiology' and 'manifestation' has been solved by adding the words "aetiology" and "manifestation" to the search text and '+' or '*' to the ICD codes.
5. Many ICD codes, especially in the malignancies and tumours, have no appropriate search text: they are included in the list of the mappings but there is no search term that refers to it. This makes that some common diagnoses (even in primary care) cannot be found by using this AT (only one example: benign tumour of the bladder).
6. The AT contains more than 12.000 "see" and "see also" referrals (these contain many 'synonyms' or 'analogous' terms) instead of direct links to the specific ICD/ICPC codes: in a actual computer system these referrals can be implemented in the search term list as such, and these search terms with synonyms and analogous terms are different for each language. So referrals are not useful in a Bilingual system, or two separate lists should have to be created.

The way of working.

Preceding remark: the way of constructing the 3BT has been 'growing on' during the process of development. By consequence of this way of working, the different steps were not always in the most logic order and a lot of time has been lost by going several times through the same list to implement changes that could have been done at once.

Successive steps (all numbers are approximate):

- From the 109.000 items we eliminated temporary the 12.000 items without link to codes but only with a link to “see” or “see also” label, referring to other existing search terms.
- With all of the 97.000 remaining items (collections of search terms or sentences) a ‘clinical label’ has been created. A ‘clinical label’ is defined as a ‘fixed concept’ or ‘unit of thought’ described in a way it becomes useful as text to implement in the EPR as assessment, using ‘natural language’ and linguistic correct medical terminology. Every clinical label at first time has been created only according the content of the search terms of the AT. Only in a later phase of the process, the created ‘clinical labels’ have been adopted to the content of the mapped ICD and/or ICPC code title, in harmony with its inclusion and exclusion criteria.
- At first it has been done for all of them (97.000), even for all the identical with different order within the search terms. At that moment it has not yet been decided to eliminate all these ‘double labels’. Even worse: we created labels with different order of words (question of implementing a bit of variation in the list of the labels), so it was even more complicated to find in a later phase all of the “equal” labels.
- Later in the developing of the ‘clinical labels’ the content of the ICD title and the content of the ICPC title has been taken into consideration in a way that the newly created label illustrate correct the content of the two codes, and in a way that many doubles (identical) are avoided. (Example: with every ICPC code about inborn errors or diseases the clinical label will contain the word ‘inborn’ if the medical ‘label’ could be used in another sense but inborn).
- ICPC chapter X and Y contain many doubles (identical labels), but to avoid the GP having a too long search list containing many ‘identical’ labels, the EPR developer has to organize it in a way that with a male patient X codes cannot be used and with female patients Y codes cannot be used and are not visualized.
- ‘Aetiology’ and ‘manifestation’ terms were added to the newly created labels, to show the different ‘meaning’ or ‘use’ of the linked codes.
- Next we corrected a lot of linguistic errors that were made (lack of good agreements at start) and eliminated as much as possible all of punctuation marks and non-common abbreviations (in the Netherlands other abbreviations and acronyms are common than in Flanders). In the 3BT only general accepted and well-known terms and acronyms and ‘letter-words’ are used.
- To change the mapping and make the 3BT ‘ICPC oriented’ all of the mappings (especially the ‘many ICPC to one ICD’) had to be reorganised and rewritten in a way the label was correct with the linked mapping of codes. Priority was given to the content of the ICPC codes and titles.
- In a later phase of the process it has been decided to eliminate all of the ‘double labels’ to create a system in which the order of the search terms has no importance.
- Once the created and selected 47.000 ‘clinical labels’ were almost ready (a process that at first has been done almost exclusively in Dutch) the French-speaking group started the work of translating the ‘clinical labels’.
- In order to avoid identical labels and to avoid the additional marker ‘aetiology’ or ‘manifestation’ added to these clinical labels, and because this concept can

be confusing, all the identical labels with the marker 'aetiology' have been eliminated. In the future it could be that the aetiology labels will be added again according to the feed back given by the users or according new solutions decided by WICC.

- Search terms at this moment (in Dutch) still contained a lot of words – originating from AT-, which are not used as keywords in the search system of an EPR. This was a next step: cleaning up all of the 'search terms' from the AT in a way it contained only sensuous keywords and no punctuation marks.
- The search system had to be improved by adding many synonyms and analogous words to the search terms. To improve this aspect of the thesaurus comments and feed back from users were very helpful.
- The process of developing search terms has been done independently by the Dutch and the French team, because the French team had no resource that had been used in the creation of the 3BT and the Dutch team had to correct and improve the search terms originating from the AT. Also adding synonyms and analogue terms is something that's specific for each language.
- The last point was looking at the old 12.000 "see" and "see also" referral list to add most of the used terms in this list to the search terms of the remaining 47.000 "clinical labels"
- Result: in Dutch there are an average of 5 terms (keywords) by clinical label and about 4 in French.
- Revision of ICPC-ICD mapping according latest official version of ICPC-2 (march 2005)
- In developing the labels some procedures included in ICD has been eliminated because of overlap with procedures in ICPC
- Procedures: independent of the AT a Belgian list of procedures has been added to the ICPC process codes (component 2-6) as extension of ICPC codes *30-*69. These procedures include everything what the GP is doing in the own practice or is demanding for in own decision making-process. It contains also all the reimbursed procedures in Belgian health care system. The process codes are not chapter linked, in conformity with ICPC-2 version, and they are not linked to ICD codes.
- Each of the 47.000 'clinical labels' has a unique identifier, which guarantees the link from Dutch and French identical clinical labels with the matching ICPC/ICD codes and with each other, and it generates the link with the appropriate keywords for each 'clinical label'.
- Continuous follow-up: adding missing concepts (bilingual), improving the list of keywords for each language separately and updating the ICPC/ICD mapping according to the mapping corrections by WICC.

Marc Verbeke
Zeile
September 26, 2006

Developing team: Ghent University
Jan De Maeseneer, Diëgo Schrans, Sven Deroose, Marc Verbeke, Paul Van Hove
Brussels University
Michel Roland, Michel Dejonghe, Nadine Kacenenbogen, Bernard Dendeau
Federal Government
Marc Bangels, Luc Nicolas

Suggestions for creating a new thesaurus in other languages.

1. There could be started from the AT English version on the CD in the book ICPC-2-R. All the doubles (see word order) has to be eliminated first, and then create clinical labels.
2. Another way would be by translating the created clinical labels of the 3BT by help of the Spanish versions of ICPC and ICD. In a second time the list of keywords has to be created by using first all of the words within the labels, words from inclusion criteria in ICPC, and adding other synonyms and analogous words. Criterion: which terms is the doctor using in searching for a concept?
3. In both ways of working first some decisions has to be taken:
 - a. Use of “see” and “see also” referrals
 - b. Use of “aetiology” and “manifestation” concept
 - c. What will be the prior access: ICPC (basic structure of 3BT) or ICD (basic structure of AT)?
 - d. What has to be done with procedures included in ICD and included in ICPC codes *30-*69?
 - e. Use of punctuation marks and numbers in the labels
 - f. Which words are senseless in use as keyword, but are part of the clinical labels
 - g. Creating a structure for linking the two languages.